Weighted likelihood in Bayesian inference

Claudio Agostinelli and Luca Greco

Abstract The occurrence of anomalous values with respect to the specified model can seriously alter the shape of the likelihood function and lead to posterior distributions far from those one would obtain without these data inadequacies. In order to deal with these hindrances, a robust approach is discussed, which allows us to obtain outliers' resistant posterior distributions with properties similar to those of a proper posterior distribution. The methodology is based on the replacement of the genuine likelihood by a weighted likelihood function in the Bayes' formula.

Key words: Likelihood uncertainty, Outliers, Robust posterior distribution, Weighted Likelihood

1 Introduction

The problem of formalizing subjective uncertainties in the probability model for the data has a central role in Bayesian robustness [2, 4]. Model uncertainty, essentially, comes from the presence of outliers, - observations that are highly unlikely to occur under the assumed model— [6], as, for instance, they obey a different and unsuspected (hence unspecified) random mechanism. Actually, because of the presence of some anomalous values, the information about the variations of statistical evidence, summarized by the likelihood function, can be seriously misleading and invalidate the updating mechanism of our initial knowledge. The application of the Bayes' formula can lead to a posterior distribution whose shape may be dramatically differ-

C. Agostinelli
Dipartimento di Scienze Ambientali, Informatica e Statistica and Scuola Superiore di Economia, Università Ca' Foscari, Venezia, Italia

L. Greco
Dipartimento degli Studi dei Sistemi Economici, Giuridici e Sociali, Università degli Studi del Sannio, Benevento, Italia

ent from that one would obtain when, for instance, no outliers are in the sample at hand. Then, it seems reasonable to look for a methodology that results in outliers insensitive posterior distributions and robust posterior summaries with respect to data inadequacies.

One recent approach to handle uncertainty about the sampling model and to obtain a robust posterior distribution has been outlined in [3]. The authors investigate the use of a quasi–likelihood function with robustness properties in place of the genuine likelihood function and prove its validity in order to perform Bayesian inference along the lines illustrated in [8, 5]. the empirical likelihood has been considered as well.

Here, a more general strategy is discussed, which is based on the replacement of the genuine likelihood by a weighted likelihood function in the Bayes' formula. The weighted likelihood is characterized by the introduction of a set of weights which are aimed at down-weighting those likelihood single term components that correspond to anomalous values in the sample at hand. Under the assumed model, the weighted likelihood shares the main (asymptotic) features of the genuine likelihood function, thus being valid for Bayesian inference. This means that, once one has assumed a sampling model and a prior distribution, by applying the Bayes' theorem this strategy allows us to obtain a proper posterior distribution, i.e. that obeys probability laws, and make reliable inference in the presence of anomalous data in the sample.

2 Weighted Likelihood

Let $\mathbf{x} = (x_1, \cdots, x_n)$ be an i.i.d. sample of size n drawn from a random variable X with unknown probability (density) function $m(x|\theta)$, which is an element of the parametric family $\mathscr{M} = \{m(x|\theta), \theta \in \Theta \subset \mathbb{R}^p\}$, with $p \geq 1$. Let $\hat{F}_n$ be the empirical cumulative distribution function based on the sample $\mathbf{x}$. A weighted likelihood function can be defined as

$$L^w(\theta) = L^w(\mathbf{x}|\theta) = \prod_{i=1}^{n} m(x_i|\theta)^{w(x_i)} \;, \tag{1}$$

where $w(\cdot)$ is a bounded differentiable non negative (weight) function that may depend on unknown quantities (say η) and/or on the random sample $\mathbf{x}$. In most cases $\theta \subseteq \eta$, but there are also cases where there is no relation between the two vectors.

Under the classical regularity assumptions on the likelihood, the properties of (1) and related quantities are driven by the behavior of the weighted score function

$$s^w(\theta) = s^w(\mathbf{x}|\theta) = \sum_{i=1}^{n} w(x_i) s(x_i|\theta) \tag{2}$$

where $s(x|\theta)$ denotes the ordinary score function.

Proposition 1. *Let $\hat{\eta}$ be a consistent estimator of the unknown parameter η and $E(s(X|\theta)^2) < \infty$. Assume that the weight function $w(x;\hat{\eta},\hat{F}_n)$ is such that*

$$\sup_x |w(x;\hat{\eta},\hat{F}_n) - c| \xrightarrow{p} 0 \qquad as \qquad n \to \infty \ ,$$

where c is a positive constant. Then

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i;\hat{\eta},\hat{F}_n)s(x_i|\theta) \xrightarrow{p} c\, E(s(X|\theta)) = 0 \ .$$

Proposition 1 provides a sufficient condition on the weight function such that the equation (2) defines an unbiased estimating equation at the assumed model as well as the ordinary score function. Similar results are also valid for other likelihood based quantities.

A weight function satisfying Proposition 1 can be obtained by following the proposal of [6] which is related to the minimum distance methods. This weight function is defined as

$$w_W(x;\theta,\hat{F}_n) = \frac{A(\delta(x;\theta,\hat{F}_n))+1}{\delta(x;\theta,\hat{F}_n)+1} \ , \tag{3}$$

where

$$\delta(x;\theta,\hat{F}_n) = \frac{f^*(x) - m^*(x|\theta)}{m^*(x|\theta)}$$

is the Pearson residual function, $A(\cdot)$ is the Residual Adjustment Function, $f^*(x)$ is a nonparametric kernel density estimate and $m^*(x|\theta)$ is a smoothed version of the model density obtained by using the same kernel function.

The weight function (3) depends on the unknown θ. As an estimate of θ, we take the root $\hat{\theta}_W$ of the Weighted Likelihood Estimating Equations (WLEE)

$$\sum_{i=1}^{n} w_W(x_i;\theta,\hat{F}_n)s(x_i|\theta) = 0 \ . \tag{4}$$

Under the assumed model (or equivalently, when no outliers occur) and classical regularity assumptions, as the sample size increases, all the weights defined in (3), with θ replaced by $\hat{\theta}_W$, tend to unity uniformly almost surely, i.e. $c = 1$. Therefore, $\hat{\theta}_W$ is a consistent and first order efficient estimator of θ, that is $\sqrt{n}(\hat{\theta}_W - \theta) \xrightarrow{d} N(0, i_1^{-1}(\theta))$ and, by Proposition 1, the resulting weighted likelihood function shares the same first order asymptotic properties of the genuine likelihood function, hence leading to estimators and tests with the usual asymptotic behavior [1].

3 Weighted posterior distributions

In Bayesian inference, the posterior distribution $\pi(\theta|\mathbf{x})$ is obtained by combining two sources of information about the random variable θ: one is the prior knowledge about its distribution $\pi(\theta)$, the other is given by the observed data and is summarized by the likelihood function. A weighted posterior distribution is defined by replacing the genuine likelihood function by its weighted counterpart defined in (1), i.e.

$$\pi^w(\theta|\mathbf{x}) \propto \pi(\theta)L^w(\mathbf{x}|\theta) \ . \tag{5}$$

In the following, we consider weights evaluated as in (3) of the form $w(x_i) = w_W(x_i; \hat{\theta}_W, \hat{F}_n)$. As $L^w(x|\theta)$ shares the first order properties of the genuine likelihood function under the assumed model, it is valid for Bayesian inference in a standard fashion.

This methodology has the great advantage, over the employ of other pseudo-likelihood functions, to lead to posterior distributions which belong to the same family of those one would obtain by using the genuine likelihood function. Hence, the weighted posterior distribution will differ from the genuine posterior distribution only for the value of its parameters.

The method may appear in conflict with a proper Bayesian perspective, since the weighted likelihood is not directly driven by a probabilistic model, but by using adaptive weights, we still make the all data tell their all story, as some values are recognized as inconsistent with the model and these outliers are not simply deleted but still contribute to posterior inference even if in a different fashion than ordinary.

3.1 Rats data

This data set [7] corresponds to an experiment on the speed of learning of a rat to go through a shuttlebox in successive attempts. If the recorded time was larger than 5 seconds, the rat received an electric shock. The entries, for each observation, are the average time for all attempts between shocks (Time) and the number of shocks received (Shocks). The model is $\mathrm{Time_i} = \beta_0 + \beta_1 \mathrm{Shocks_i} + \sigma u_i$, $i = 1, 2, \ldots, 16$ and the u_i's are i.i.d. standard normal variates. Because of the presence of three points which, apparently, lie far from the others, a robust regression by weighted likelihood [6], with a RAF based on the Hellinger distance, seems appropriate. The robust fit strongly down–weights only observation n. 4. The weighted joint posterior distribution is

$$\pi^w(\beta, \sigma^2|y) \propto \sigma^{-\sum_{i=1}^n w_i - 2} \exp\left\{ -\frac{(\sum_{i=1}^n w_i - 2)s^2}{2\sigma^2} \right\} \exp\left\{ -\frac{1}{2}(\beta - \hat{\beta}_W)^T V (\beta - \hat{\beta}_W) \right\}$$

where $s^2 = (\sum_{i=1}^n w_i - 2)^{-1}(\beta - \hat{\beta}_W)^T W(\beta - \hat{\beta}_W)$,with $W = diag(w_i)$ and

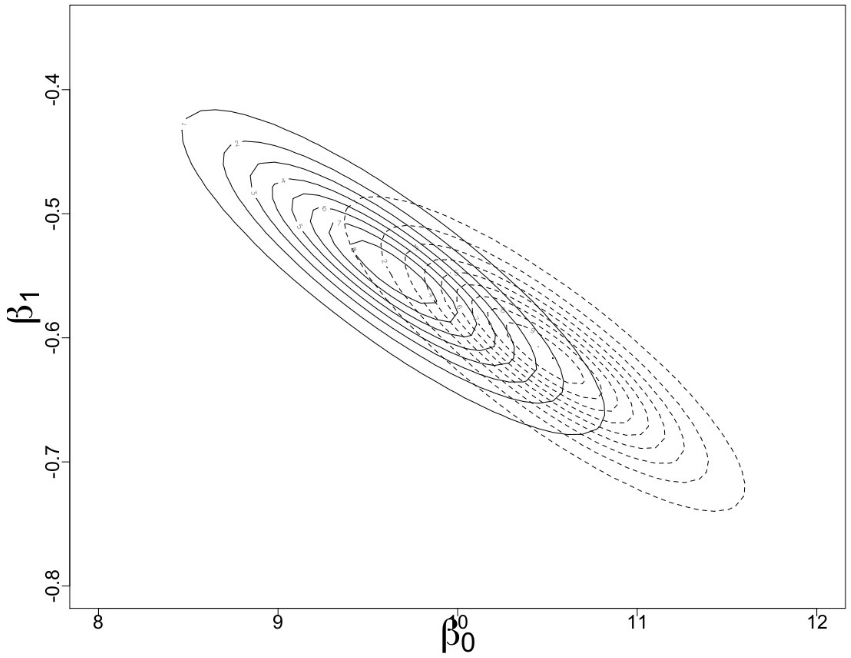

Fig. 1 Rats data. Contours of the marginal posterior density of (β_0, β_1) by WLEE–weights (solid line) and of the proper posterior density (dashed line).

$$V = \begin{pmatrix} \sum_{i=1}^{n} w_i & \sum_{i=1}^{n} x_i w_i \\ \sum_{i=1}^{n} x_i w_i & \sum_{i=1}^{n} x_i^2 w_i \end{pmatrix} .$$

The weighted marginal posterior densities of (β_0, β_1) and σ^2 are displayed in Figure 1 and Figure 2, respectively, for a Jeffreys' prior distribution. The weighted marginal $\pi^w(\beta|y)$ is multivariate Student with $\sum_{i=1}^{n} w_i - 2$ degrees of freedom, whereas $\pi^w(\sigma^2|y)$ is inverse gamma. The main effect of the weights is that of shifting the marginal posterior distributions over different regions of the parameter space, so resulting in credible sets different from those one would have obtained according to the misleading information driven by the presence of the outlier.

3.2 Poisson regression

The use of a set of weights and the replacement of the genuine likelihood function by its weighted counterpart is not supposed to compromise the possibility to face more complex problems in which it is custom to turn to MCMC algorithms. The weighted

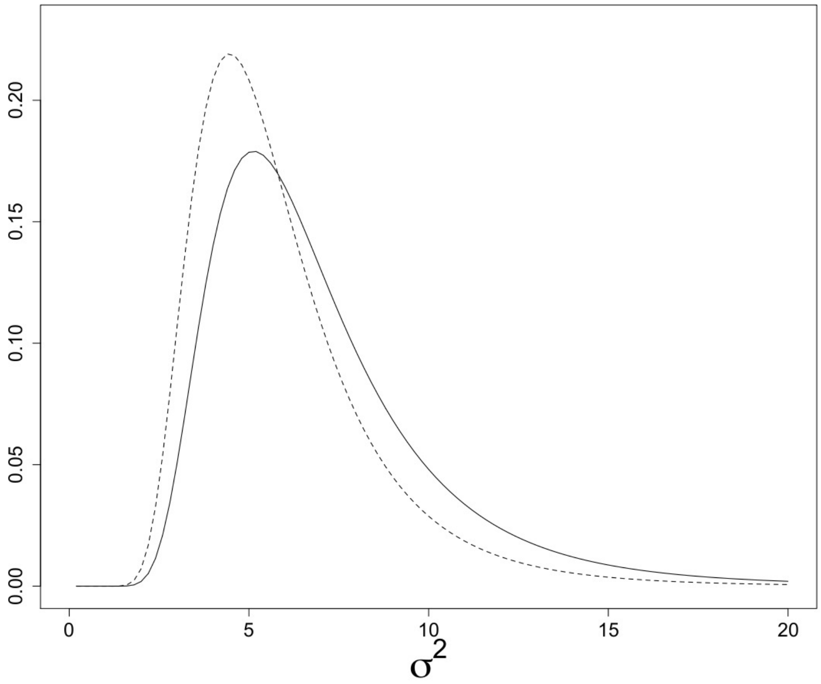

Fig. 2 Rats data. Marginal posterior density of σ^2 by WLEE–weights (solid line) and of the proper posterior density (dashed line).

likelihood can be employed for Bayesian inference in a Poisson generalized linear model of the form

$$Y_i \sim Pois(\mu_i), \quad \log \mu_i = x_i^T \beta, \quad \beta = (\beta_0, \beta_1, \dots, \beta_4)^T, \quad i = 1, 2, \dots, n,$$

with a multivariate normal prior distribution $\pi(\beta)$ with zero mean vector, uncorrelated components and large diagonal values, in order to use a diffuse prior. The latter has been preferred with respect to an improper uniform prior for computational reasons, in particular to allow the evaluation of the Laplace approximation in model comparison through Bayes Factors. The weighted likelihood function is obtained by using WLEE–weights with a RAF based on the Generalized Kullback-Leibler divergence.

A random walk Metropolis algorithm has been implemented to simulate from the weighted posterior distribution in a very standard fashion. In particular, the proposal distribution involved in the weighted MCMC algorithm is centered at the current value of β and has a variance-covariance matrix that depends on the large sample variance-covariance matrix of the weighted likelihood estimator $\hat{\beta}_W$ of β. The statistical environment R was used to perform all the computations. The weighted MCMC routines were implemented in R calling suitable C++ functions. An R-package is currently under development and routines are available from the authors.

Suppose that we aim at comparing the reduced model m_R indexed by $\beta_R = (\beta_0, \beta_1, \beta_2)$, with the full model m_F, indexed by β. To this end a weighted Bayes Factor (WBF) is properly defined as

$$BF^{w}_{m_R/m_F} = \frac{\int_{\mathbb{R}^3} L^w(x|\beta_R)\pi(\beta_R)d\beta_R}{\int_{\mathbb{R}^5} L^w(x|\beta)\pi(\beta)d\beta} \ , \tag{6}$$

where $w(y_i, x_i^T) = w_W(y_i, x_i^T; \hat{\beta}_W, \hat{F}_n)$ are evaluated only only under m_F.

References

1. Agostinelli, C. and Markatou, M.: Test of Hypothesis based on the Weighted Likelihood Methodology. St.Sinica **1**, 499–514 (2001)
2. Berger, J.O. : An overview of robust Bayesian analysis (with discussion). Test **3**, 5–124 (1994)
3. Greco, L., Racugno, W., Ventura, L.: Robust Likelihood functions in Bayesian analysis. J.St.Pl.Inf. **138**, 1258–1270 (2008)
4. Huber, P., Ronchetti, E.: Robust Statistics (second edition). John Wiley & Sons, Inc., Hoboken, New Jersey (2009)
5. Lazar, N.A.: Bayesian empirical likelihood. Biomtrka **90**, 319–326, (2003)
6. Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood estimating equations with a bootstrap root search. JASA **93**, 740–750 (1998)
7. Maronna, A.R., Martin, R.D., Yohai, V.J.: Robust Statistics. Theory and Methods. John Wiley & Sons Ltd, Chichester (2006)
8. Monahan, J.F., Boos, D.D.: Proper likelihoods for Bayesian analysis. Biomtrka **79**, 271–278 (1992)